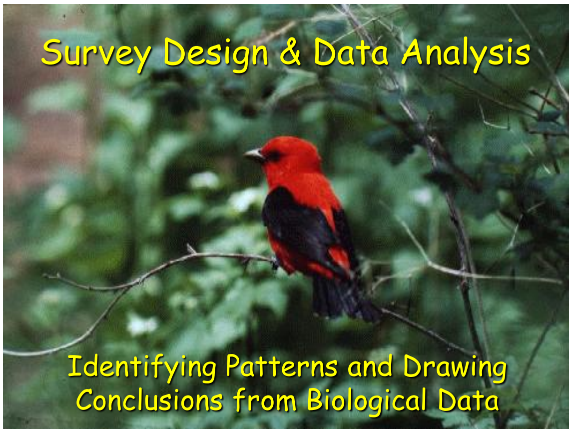


Take Home Lessons

Statistics are used to simplify patterns underlying complex biological phenomena

Consultation with a statistician should be mandatory for any survey-based project

Take Home Lessons

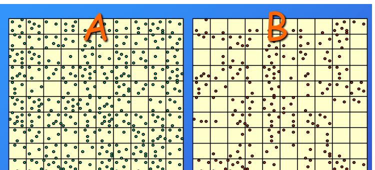
Statistics are used to support the decision-making process, and not intended to be the sole consideration in that process

Strong experimental design and statistical analyses lend irrefutable credibility to survey results

Do we need statistics?

Census

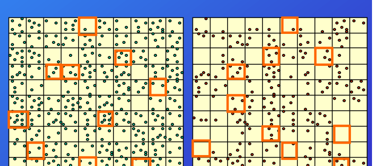
Conclusion:
Site A has a higher density of mahogany than Site B



400 trees/100 ha   235 trees/100 ha

Statistics
(using sampling)

Conclusion:
There is a 95% chance that Site A has a higher density than Site B



4.1 $\pm$ 0.5 trees/10 ha   2.4 $\pm$ 0.6 trees/10 ha

Statistics:

Descriptive vs Inferential

Describes a sample of the population

Uses sample to generalize to entire population

Major Steps in designing, Implementing and Evaluating a Project

1. What is my question or hypothesis?
2. What parameters need to be estimated?
3. Can the parameter be reliably estimated?
4. How will the project be designed?
5. How will the data be analyzed?
6. How will the project be evaluated?

1. What is the Question or Hypothesis?

The most important step

What

Where

When

Sampling Universe

The population about which you want to draw conclusions

Birds migrating through the Westwoods National Monument

Question: Does the abundance of neotropical migratory birds differ among forest interior and young forest/edge habitats at the Westwoods National Monument during Spring migration ?

H_o: Abundance of NTMBs is the same in both habitat types

H_A: Abundance of NTMBs is not the same in both habitat types

2. What Parameter Needs to be measured?

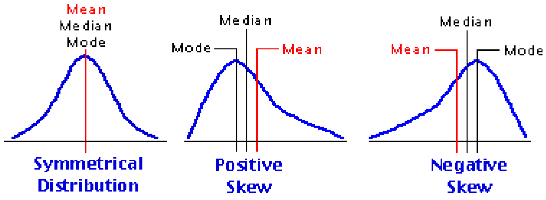
Quantitative characteristic of a population

- Number of individuals
- "Health" of the population
- Environmental threat
- Other characteristics of population

Types of Biological Data

- Nominal -- Attribute rather than quantitative
  Male, female   blue, red, green
- Ordinal -- Relative difference or ranking
  Small, medium, large   A, B, C, D,...
- Discrete -- Quantitative, only whole numbers
  0, 1, 2, 3, 4,...
- Continuous -- Any whole or mixed number
  3.4, 19.67, 12.975,...

Ecological data often are not taken from a symmetrical, bell-shaped curve



3. Can the Parameter be Measured Reliably?

Can you collect enough data to address the question of interest?

Are you measuring the population that you said you would measure?

Is there excessive ERROR or BIAS in your measurements?

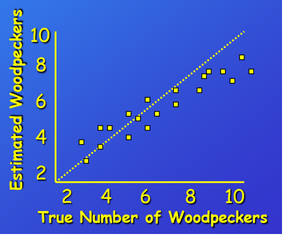
Sampling

Error

Random deviations from the true values

Bias

Systematic deviations from the true values



Major Steps in designing, Implementing and Evaluating a Project

1. What is my question or hypothesis?
2. What parameters need to be estimated?
3. Can the parameter be reliably estimated?
4. How will the project be designed?
5. How will the data be analyzed?
6. How will the project be evaluated?

4. What is an Appropriate Study Design?

First, go back to Step 2 to review what biological characteristic you are trying to measure

➡ Overall abundance of NTMBs

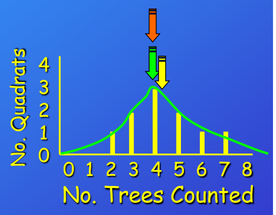
Second, determine what statistical parameter you need to estimate

➡ Mean number of NTMBs per plot

Measures of Central Tendency

What is most typical for this population

Mean -- average
Median -- middle
Mode -- most common

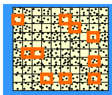


Normal Distribution -- bell-shaped curve

Measures of Central Tendency

Most often the MEAN is used as the parameter of interest

But, the value of the mean tells us little without an indication of the DISPERSION of values used to calculate that mean.



Mean = 4 + 2 + 5 + 4 + 7 + 3 + 5 + 3 + 6 + 4 = 4.3

Variance (s^2) & Standard Deviation (sd)

$$S^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = \frac{205 - \frac{(43)^2}{10}}{10-1} = \frac{205 - 184.9}{9}$$

= 20.1 / 9 = 2.23 = Variance

x	x^2
4	16
2	4
5	25
4	16
7	49
3	9
5	25
3	9
6	36
4	16

n = 10
Σx = 43
Σx^2 = 205

S^2 = variance
x = each observation
n = no. of observations (sample size)

Standard Deviation = sd = Sqrt (Variance) = Sqrt (2.23) = **1.49**

Standard Deviation (sd)

What does it mean?

If data follow a normal distribution, then:

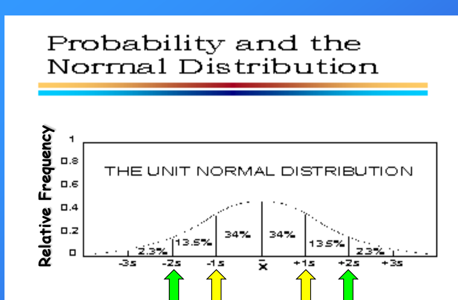
68% of all measurements are within $\pm$ 1 sd

95% of all measurements are within $\pm$ 2 sd

$\bar{x}$ = 4.30 trees/10 ha   sd = 1.49

68% of data are between 4.30 $\pm$ 1.49 (2.81 - 5.79)

95% of data are between 4.30 $\pm$ 2.98 (1.32 - 7.28)



$\pm$ 1 sd

$\pm$ 2 sd

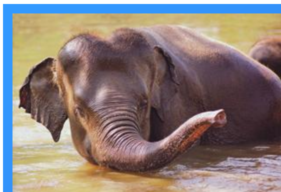
Standard Deviation (sd)
vs
Standard Error (SE)

sd = variation in the population (sample)

SE = how close estimated mean is to the true population mean

SE = sd / $\sqrt{n}$

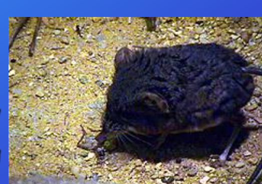
The Standard Deviation is good for examining variation around the mean, but what if we want to compare the variation in two populations that differ widely in their mean values?



Does one species show greater variation in weight?

Asian Elephant

x= 3960 kg
sd = 283



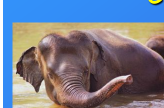
Elephant Shrew

x = 0.22 kg
sd = 0.10

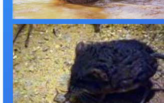
Coefficient of Variation CV)

Provides a measure of relative variability

$CV = (sd / \bar{x})(100\%)$



(283/3960)(100%) = 7%



(0.10/0.22)(100%) = 5%

Standard Deviation (sd) vs Standard Error (SE) vs Coefficient of Variation (CV)

sd = variation in the population (sample)

SE = how close estimated mean is to the true population mean = sd / sqrt(n)

CV = relative estimate of population variability = sd / mean

4. What is an Appropriate Study Design?

- First, go back to Step 2 to review what biological characteristic you are trying to measure
- Second, determine what statistical parameter you need to estimate

Third, develop sampling protocol that allows reliable data to be collected

What will be our sampling protocol?

Line transects?

Mist netting?

Spot mapping?

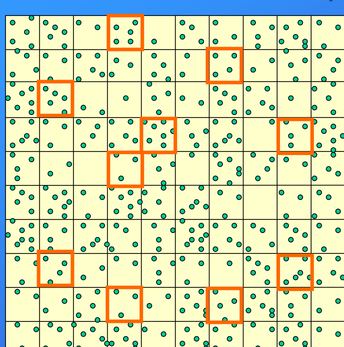
Point counts?

Sampling Design

Randomness -- Every unit in the population has an equal chance of being sampled.

Independence -- Knowing something about one unit doesn't provide information about another unit (or one unit does not influence another unit).

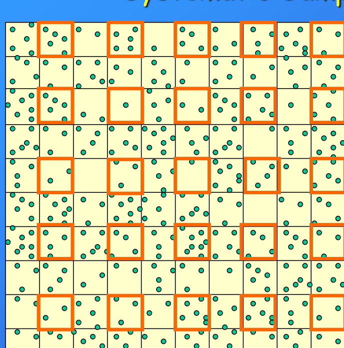
Random (or Probability) Sampling



Quadrats selected totally by chance

Random numbers table

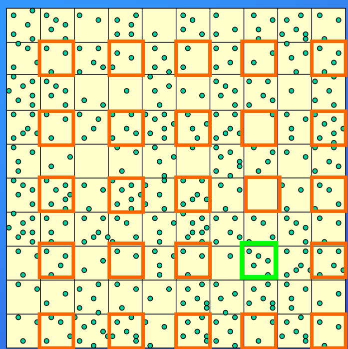
Systematic Sampling

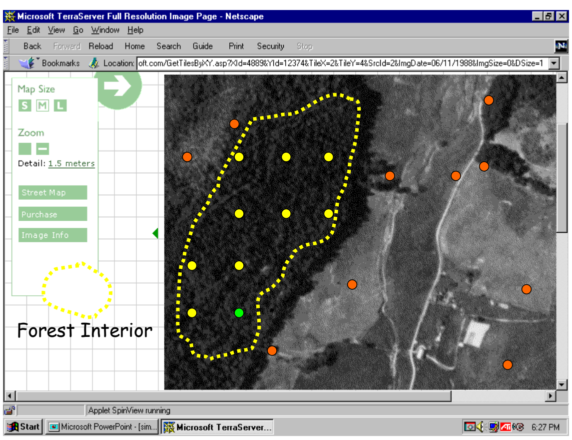


Quadrats selected are evenly spaced

No random component

Systematic Sampling with Random Start





Minimum Sample Size Requirements

or

How do we know how many point count stations we might need?

Minimum Sample Size Requirements

Calculated BEFORE study is implemented (IF POSSIBLE)

Use PILOT STUDY or existing data

Need 4 pieces of information:

1. Mean & measure of variance of parameter
2. Magnitude of difference you want to detect
3. Significance level (α)
4. Desired power

Some Statistical Terminology

- Type I statistical error
- Type II statistical error
- Power

Statistical Errors in Hypothesis Testing

Alpha (α) -- Also called Type I error. Probability that we reject the Null Hypothesis when in fact it is true.

Beta (β) -- Also called Type II error. Probability that we do not reject the Null Hypothesis when it is, in fact, false.

The Power of Statistical Tests

The probability of rejecting the Null Hypothesis when, in fact, it is false (and should be rejected)

Power = 1 - β

Important for calculating minimum necessary sample sizes

Minimum Sample Size Requirements

Calculated BEFORE study is implemented (IF POSSIBLE)

Use PILOT STUDY or existing data

Need 4 pieces of information:

1. Mean & measure of variance of parameter
2. Magnitude of difference you want to detect
3. Significance level (α)
4. Desired power

Minimum Sample Size (n)

$$n = \frac{2Ms^2}{d^2}$$

Where, d = minimum detectable difference
M = multiplier from normal distribution
s^2 = estimated population variance

5. Appropriate Data Analysis

Good analyses begin
with good hypotheses

All statistical tests must have
two types of hypotheses:

Null Hypothesis (H_o)
Alternative Hypothesis (H_A)

Null Hypothesis usually is
tested via a statistical test

If Null Hypothesis not accepted,
then Alternative Hypothesis
is assumed to be true

We need an objective way of rejecting or not rejecting the null hypothesis based upon the probability that the estimated parameter occurred by chance alone.

The conclusion that the observed result is significant is established by the significance level (alpha or α).

It is the probability above which we do not reject the Null Hypothesis.

Now, let's get down to business…

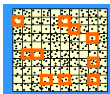
Question: Does the abundance of neotropical migratory birds differ among forest interior and young forest/edge habitats at the Westwoods National Monument during Spring migration?

H_o: Abundance of NTMBs is the same in both habitat types (α = 0.05)

H_A: Abundance of NTMBs is not the same in both habitat types

Let's take a quick look at the data…

Plot	Forest Interior		Young Forest/Edge	
1	3		4	
2	7		5	
3	4		5	
4	3		5	
5	2		4	
6	1		2	
7	4		3	
8	5		6	
9	3	$\bar{x}$ = 3.60	6	$\bar{x}$ = 4.90
10	4	s = 1.65	9	s = 1.91



FOREST INTERIOR

Mean = 3 + 7 + 4 + 3 + 2 + 1 + 4 + 5 + 3 + 4 = 3.60

Variance (s^2) & Standard Deviation (sd)

$$S^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = 2.72 = \text{Variance}$$

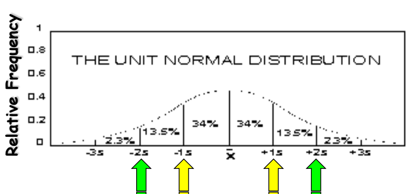
x	x^2
3	9
7	49
4	16
3	9
2	4
1	1
4	16
5	25
3	9
4	16

n = 10
Σ x = 36
Σ x^2 = 154

S^2 = variance
x = each observation
n = no. of observations (sample size)

Standard Deviation =
sd = Sqrt (Variance) =
Sqrt (2.72) = 1.65

Probability and the Normal Distribution



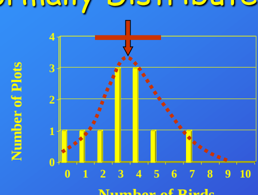
$\pm$ 1 sd ---------> 68%

$\pm$ 2 sd ---------> 95%

Are the Data Normally Distributed?

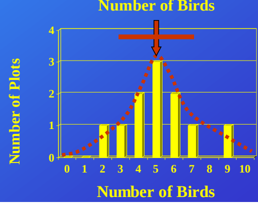
Forest Interior

$\bar{x}$ = 3.60
s = 1.65



Young Forest /Edge

$\bar{x}$ = 4.90
s = 1.91



So…
The data appear to follow an (approximate) normal distribution

So…
We can use parametric statistics to analyze the data

We choose a t-test, because:

We are comparing only two samples

The data are normally distributed

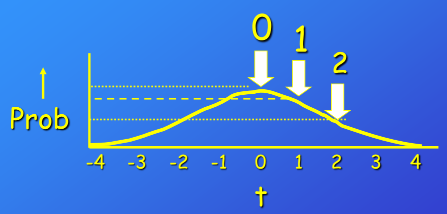
The two samples have similar variances

The t-test will allow us to assess if there is a difference in the abundance of NTMBs in the two habitat types, Forest Interior and Young Forest/Edge

$$t = \frac{\bar{X} - \bar{X}}{\sqrt{\frac{\left(\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - 1}\right)}{n} + \frac{\left(\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - 1}\right)}{n}}}$$

The test statistic…

$t = 1.63$



Look up critical value in table: α = 0.05 (2-tailed), v = 18

What do we conclude?

H_o: Abundance of NTMBs is the same in both habitat types

H_A: Abundance of NTMBs is not the same in both habitat types

Since the t-statistic (1.63) is not greater than the critical value from the table (2.101), we cannot reject the null hypothesis and conclude that no difference exists between habitats.

Another way to examine the differences between the mean values of two samples

Confidence Intervals

An estimated range of values which is likely to include an unknown population parameter

Often, 95%

Allows us to estimate the precision of our estimate

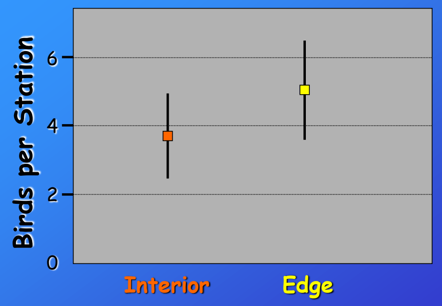
$$\bar{x} \pm (t_{\alpha(2),n-1}) \left(\frac{sd}{\sqrt{n}} \right)$$

t value taken from statistical table

Estimated mean

Standard deviation divided by square root of sample size

Mean + 95% Confidence Intervals



Now, back to our hypothess testing

What do we conclude?

H_o: Abundance of NTMBs is the same in both habitat types

H_A: Abundance of NTMBs is not the same in both habitat types

Since the t-statistic (1.63) is not greater than the critical value form the table (2.101), we cannot reject the null hypothesis and conclude that no difference exists between habitats.

What do we conclude?

H_o: Abundance of NTMBs is the same in both habitat types

H_A: Abundance of NTMBs is not the same in both habitat types

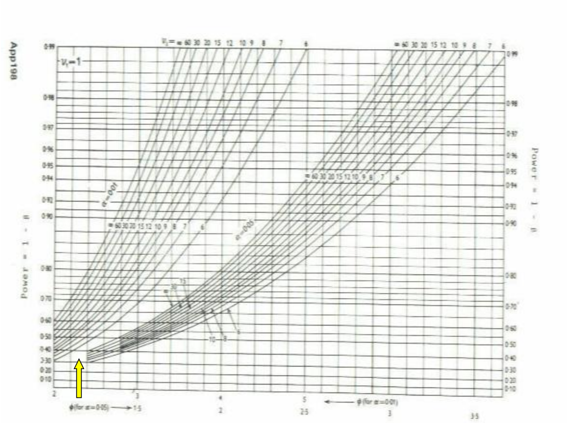
Since the t-statistic (1.63) is not greater than the critical value form the table (2.101), we cannot reject the null hypothesis and conclude that no difference exists between habitats.

But, was our design rigorous enough to be able to detect a difference?

In other words, what was our Power and Type II error?

$$\phi = \sqrt{\frac{n\,(\text{difference}^2)}{4\,(\text{variance})}}$$

$= 1.15$ Look up Power in Figure B.1a



The Power of this statistical test was approximately 0.30, which means…

That there was only a 30% chance that we could have detected this difference

or

that there was a 70% chance that we claimed there was no difference in NTMB abundance when, in fact, there was a difference

How many point counts would we have needed in each habitat to detect a 26% difference?

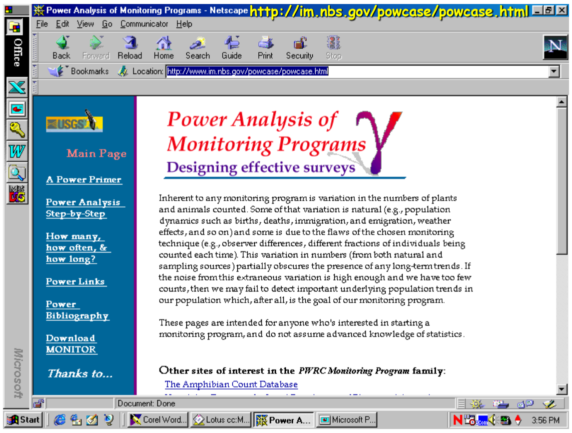
$$\frac{4.90 - 3.60}{4.90} \qquad n = \frac{2Ms^2}{d^2}$$

M = multiplier from table = 7.9

s^2 = variance = 3.648

d = difference = 4.90 - 3.60 = 1.30

$$n = \frac{2\ (7.9)\ (3.648)}{(1.30)(1.30)} = \frac{57.64}{1.69} = 34.1 \text{ points}$$



Conclusions

No significant difference existed in abundance of NTMBs between forest interior and young forest/edge habitats…

but

we had only a 30% chance of detecting a difference if it did, indeed, exist

6. Evaluating Success of Project

Completeness of data

Accuracy of data

Appropriateness of data

"Adaptive" Approach

Data Management

All survey data should be:

Checked for errors before & after recording in an electronic format

Recorded in an electronic format ASAP

Stored in two separate locations

Accompanied by metadata

Choosing an Appropriate Statistical Test

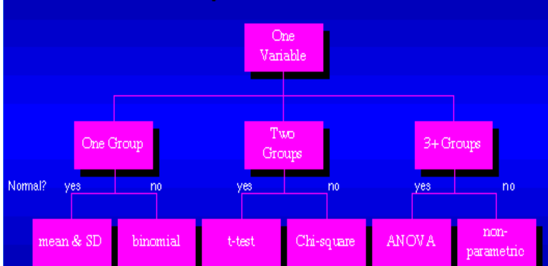
Type of Data

Number of Variables

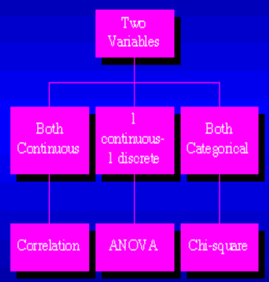
Sample Characteristics

Nature of hypothesis/research question

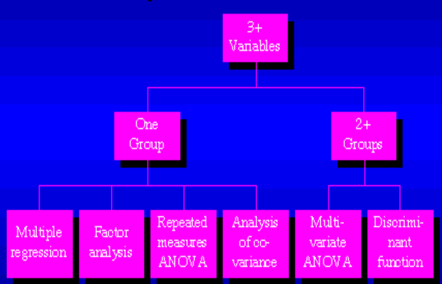
How many variables?



How many variables?



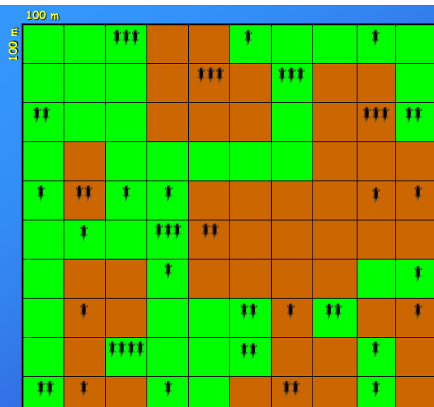
How many variables?



Example

Lizards…

…of the Venezuelan Savanna



We are interested in detecting:

A change in lizard population density of at least 50%

At a significance (α) level of 0.10

And power of 90%

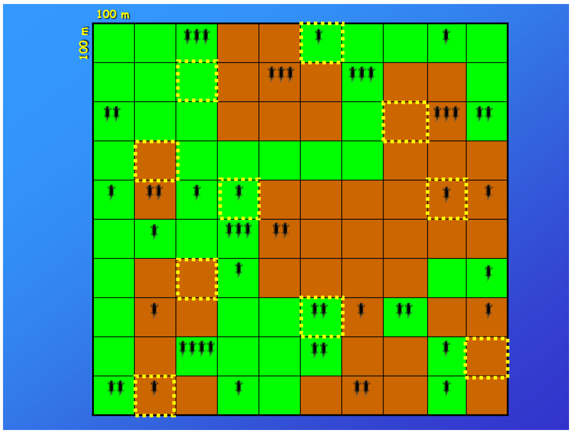
Minimum Sample Size (n)

$$n = \frac{2Ms^2}{d^2}$$

Where, d = minimum detectable difference

M = multiplier from normal distribution

s^2 = estimated population variance



Minimum sample size, $n = \frac{2Ms^2}{d^2}$

M = 8.6 Mean = 0.60 s^2 = 0.49

d = (0.50)(0.60) = 0.30

$$n = \frac{2(8.6)(0.49)}{(0.30)^2} = \frac{8.43}{0.09} = 94$$

CORRECTED Minimum Sample Size (n')

Because we plan to sample such a large proportion of total area

CORRECTED Minimum Sample Size (n')

$$n' = \frac{n}{(1 + [n/N])}$$

n' = corrected sample size

n = original sample size

N = total possible sample size

$$n' = \frac{94}{(1 + [94/100])} = \frac{94}{1.94} = \mathbf{48}$$

We're also interested in habitat use by lizards in Llanos National Park

Question: Do lizards exhibit a habitat preference in Llanos National Park?

H_o: No difference in lizard abundance between habitats

H_A: Lizards not distributed equally between habitats

Alpha = 0.05

Use Mann-Whitney U-Test to test the null hypothesis

How to calculate U

Step 1. Rank order all counts of lizards from each of the two habitats

Step 2. Sum the ranks from the smaller sample. This gives R_1. [595.5]

Step 3. Calculate U_1 from the equation:

$$U_1 = (n_1)(n_2) + \frac{n_1(n_1+1)}{2} - R_1$$

Where,

n_1 = sample size for sample 1 [21]

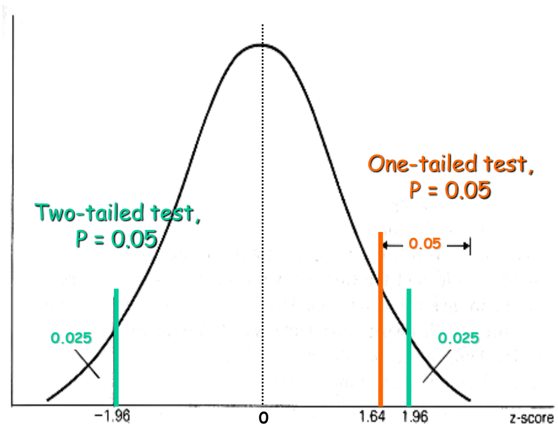
n_2 = sample size for sample 2 [27]

$U_1 = 202.5$

Step 4. Calculate U_2 from the equation:

$$U_2 = (n_1)(n_2) - U_1 \quad [U_2 = 364.5]$$

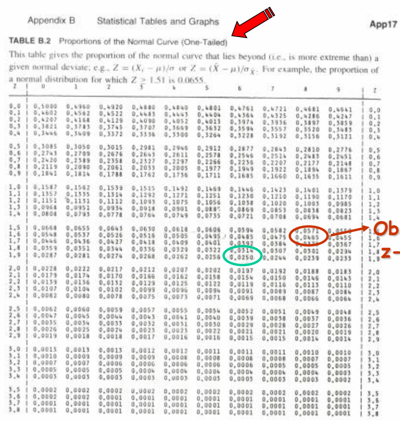
Step 5. Take the larger of U_1 & U_2 and call that U. With small sample sizes, you can compare U to values in a statistical table. But, with large sample sizes, the hypothesis must be tested using a normal approximation.



So, let's calculate our Z score
And see where it falls along the
normal distribution curve

$$Z = \frac{U - \mu_U}{\sigma_U}, \text{ where} \quad [Z = 1.68]$$

$$\mu_U = \frac{(n_1)(n_2)}{2} \qquad \sigma_U = \sqrt{\frac{(n_1)(n_2)(N+1)}{12}}$$



Appendix B Statistical Tables and Graphs App17

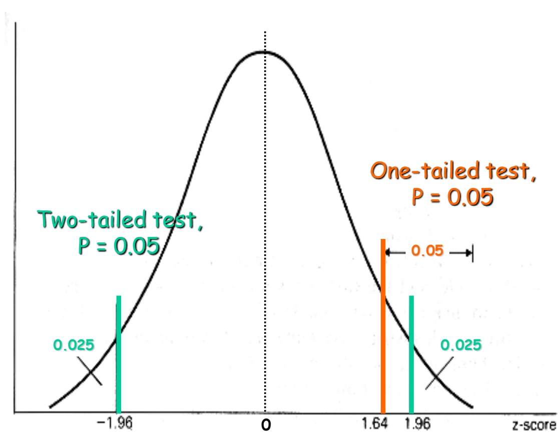
TABLE B.2 Proportions of the Normal Curve (One-Tailed)

This table gives the proportion of the normal curve that lies beyond (i.e., is more extreme than) a given normal deviate; e.g., $Z = (X_i - \mu)/\sigma$ or $Z = (\bar{X} - \mu)/\sigma_{\bar{X}}$. For example, the proportion of a normal distribution for which $Z \geq 1.51$ is 0.0655.

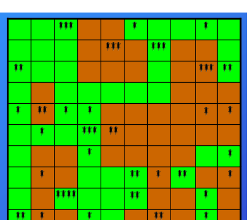
Z	0	1	2	3	4	5	6	7	8	9	Z
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641	0.0
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247	0.1
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859	0.2
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483	0.3
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121	0.4
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776	0.5
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451	0.6
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148	0.7
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867	0.8
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611	0.9
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379	1.0
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170	1.1
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985	1.2
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	[illegible]	0.0869	0.0853	0.0838	0.0823	1.3
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681	1.4
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	[illegible]	1.5
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455	1.6
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	[illegible]	0.0367	1.7
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294	1.8
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233	1.9
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183	2.0
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143	2.1
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110	2.2
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084	2.3
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064	2.4
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048	2.5
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036	2.6
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	2.7
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019	2.8
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	2.9
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010	3.0
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	3.1
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	3.2
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	3.3
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	3.4
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	3.5
3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	3.6
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	3.7
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	3.8

Z score must be $\geq$ 1.96

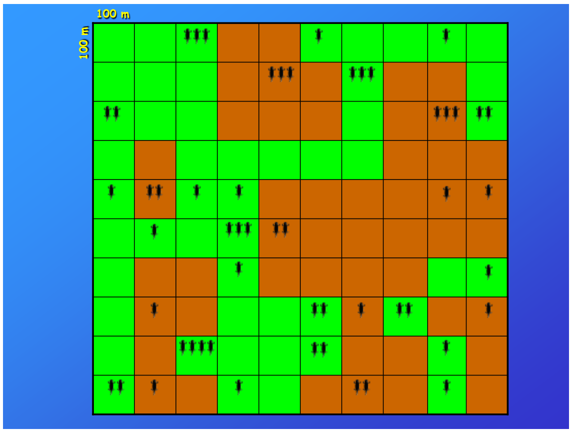
Observed z-score



What do we conclude about the abundance of lizards in the two habitat types?



How might you better design this study?


100 m
100 m